**Problem 1.1**

1. Explain why data is becoming cheaper and cheaper.

Data requires devices to store and capture it. Storage is becoming increasingly cheaper because of technology; every smart phone and computer stores data, and data centers are now abundant. There are billions of devices (some 30 billion IoT devices) capturing data and storing data, and transferring the data can be done almost immediately. Before computers were widespread, data was kept on paper and had to be collected manually, and it took more time to alter and analyze it.

2. Probability simplifies the data. Instead of looking at millions of data points, one equation (or a few equations) can be used to describe a dataset. It makes it possible for humans to understand data.

3. Data in my daily life can usually be quantified by duration, quantity of actions, type, or over a scale. Sometimes a single action can be represented by many different types of data. Like type of meals eaten: quantity of food, rating of healthiness, types of food. Others require many different data points like route driven in a car: duration of trip, duration spent in traffic, number of traffic lights, distance traveled, etc. But all those data points could simply be defined by one label.

* Number of times checked smart phone (quantity)
* time spent on smartphone (time duration)
* intensity of daily neck pain (scale)
* bathroom usage (quantity)
* Types of meals eaten (label/scale/quantity)
* Time meals eaten (time)
* Hours sitting (time duration)
* Hunger level (scale)
* Time working (time duration)
* Time relaxing (time duration)
* Type of relaxation activity (label)
* distance driven (quantity)
* route taken (label)
* Times drove (quantity)

4. Data-driven requires more resources to implement but tends to be more accurate. DD is easily able to handle large number of inputs that could not possibly be described by one equation. Instead of inferring what the result is going to be from a physics-based equation, a processor can compare input data with historical data to identify a highly likely output.

|  |  |  |
| --- | --- | --- |
| **Comparison** | **Physics** | **Data Driven** |
| human comprehension | Easy to understand relationship between inputs and outputs | Often too difficult to understand correlation between inputs and outputs |
| Data required | None, just equation | The more the better for accuracy |
| Storage required | None | Lots |
| Dimensionality | Difficult to expand past 1 or 2 dimensions. | increase dimensions easily. |
| Input Variables | Only variables in the equation can be used to estimate the output | Variables outside of what is normally considered can be added to input set to increase accuracy of estimation. |
| Variable Selection | Use dominant variable easily determined | Needs work to determine best set of variables to use |
| Output result | Estimation based on equation | Based on other results. |
| Set up required | Little, just need the law | Requires a processor, memory, data, algorithm. More resources are required to implement a data-driven approach |

**Problem 1.3**

References:

https://skymind.ai/wiki/ai-vs-machine-learning-vs-deep-learning https://www.kdnuggets.com/2015/01/deep-learning-explanation-what-how-why.html

1.

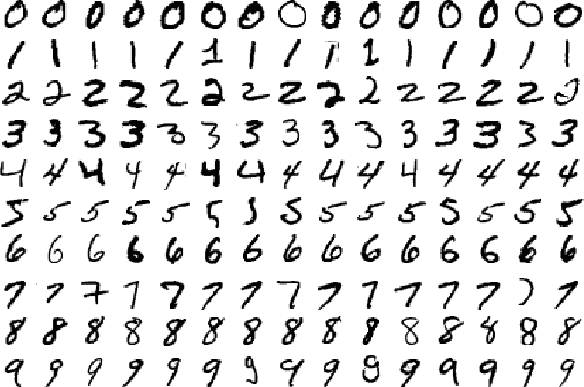
Artificial Intelligence: A program that can solve a problem in a “smart” or human way. Does not necessarily “learn”. For example, a chess program that uses a rules-based system that humans designed. Pretty much any time you play “against the computer” in a computer game, and you can choose the difficultly level of the computer you are playing against, that is traditional AI, but not machine learning. The computer can mimic a human opponent, but the boundary conditions have been pre-determined by humans. That being said, machine learning is a subset of AI.

Machine Learning: A program that can determine the set of rules based on the data that they are exposed to. So, one machine learning algorithm can learn to play checkers, and the same algorithm can learn to play chess.

Machine learning is data-driven, so it is very good at determining outliers in data. This can be useful for identity theft: normal buying patterns for credit card are groceries and gasoline in a desert, and then suddenly a jet ski is purchased. It can also be used to diagnose a sick patient based on a wide variety of symptoms, based on other patient’s diagnosis and symptoms even if it is not obvious.

Deep Learning: A subset of Machine Learning. Deep is referring to layers in a neural network. So, two or more inputs can be combined to form a more complicated set of inputs, which would be a deeper layer. For example, two points inputs can be combined to form a line, and the line would be a deeper level of the inputs that the program could use to determine the output. Currently, Deep Learning is pretty much reserved for image identification and speech recognition. A huge advantage to deep learning is that is can do feature extraction by itself, but the more features you give it, the more computational time and energy it will take.

One example of a program used by deep learning is identifying hand written notes. Determining the features to highlight for each letter or digit written is almost impossible. Differences between, 1 and 7, 5 and 3, 9 and 4 are complicated and require multiple layers of inputs to distinguish.



2.

Supervised Learning:

Machine learning that has a set of training data to learn the expected outputs based on the inputs. The goal is to provide a classification (label) or determine the expected result (regression) of an output based on the inputs. An example I heard about recently on a podcast was using pictures to identify defects in a road. You can give a supervised machine learning program a bunch of imagines of a road that has defects. Then you can take images of other roads and determine if those roads are like the defect or non-defect roads. It can then label the roads as “needs maintenance” or “not defective”.

Unsupervised Learning:

Does not need training data, and is mostly used for clustering. This method is very good at identifying anomalies in data. For example, fault detection in a power system. 99% of the time, the system is operating under normal conditions, then a fault occurs and sends all the signals out of whack. An unsupervised learning program would be able to isolate that event.

Reinforcement Learning:

A balance between unsupervised and supervised learning. Does not need training data, but the reinforcement aspect is taken from user confirmation. Reinforcement learning is used when Hulu asks the user if the Advertisement they just watched was relevant to them, or a word predictor in a texting app. Both of those applications can work without user reinforcement, but with user feedback (answering yes or no to the advertising question, or selecting a word in the texting app) the applications can improve to provide a better answer.

3.

EHarmony Dating: A compatibility score is assigned to your matches. This is done by comparing the weight of importance users assign to the 29 available inputs to determine how likely they are going to be compatible with someone else. This seems like a regression problem, or a supervised learning problem. So, basically when you sign up for Eharmony, your scores create your unique “machine learning equation”, this can now be compared with other users, after you filter out location, and sexually preferences (more filtering based on scores can be done as well to decrease time to sort matches). The program can then compare how close your features match to come up with a score.

Amazon Retail: Amazon has a massive database of products being sold. When you type in a couple key words, it picks the best options based on those. Basically it ranks the items by giving it a relevancy score. The words are not the only inputs, the review score, number of reviews, number of times purchased, price, brand, ect. By running the inputs through a regression supervised learning algorithm, the products can be ranked and listed in order of relevancy. To reduce the dataset, products without the key words in the title can be filtered out.

Page Rank: (reference: https://www.youtube.com/watch?v=LVV\_93mBfSU)

Creates a web of the internet using a spider programs, that jumps through internet sites to collect information about each website. This basically is populating the inputs that the google page rank program is going to use later to determine the page rank. Once you give google the key words, it can filter the database to only rank the websites with the words in them. The rank can then be determined weighting different inputs (another regression problem). Search terms in the title, popularity of website, average time spent on webpage, and how many times webpage is referenced by other webpages, meaning of the key words.

FitBit Wearables: FitBit and google teamed up to start using wearables data to combine with medical records to provide better recommendations on fitness. By combining medical records and daily exercise habits (resting heart rate, max heart rate, daily exercise) the inputs can be analyzed to come up with a classification (healthy, or most likely sickness) of the person, using supervised machine learning.

Pandora Music: Pandora collects 450 attributes from every song to describe a song. It also collects data, of time the song is listened to, whether the song was skipped, ect. But it also has a “thumbs up” feature, which can be considered reinforcement learning. Essentially, it is another regression problem, based on the music you currently listen to, how much does that match other music you haven’t heard yet? If a thumbs up is given to the new music, that reinforces the decision the algorithm made.